

BBN Report No. 3122

September 1975

COMMAND AND CONTROL RELATED COMPUTER TECHNOLOGY

Part I. Packet Radio

Quarterly Progress Report No. 3

1 June 1975 to 31 August 1975

The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied of the Defense Advanced Research Projects Agency or the United States Government.

The research was supported by the Defense Advanced Research Projects Agency under ARPA Order No. 2935. Contract No. MDA903-75-C-0180.

Distribution of this document is unlimited. It may be released to the Clearinghouse Department of Commerce for sale to the general public.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE(When Data Entered)

19. optimal parameter interpolation
real time signal processing
multidimensional scaling

SECURITY CLASSIFICATION OF THIS PAGE(When Data Entered)

TABLE OF CONTENTS

	<u>Page</u>
I. INTRODUCTION	1
II. MEETINGS AND PUBLICATIONS.	3
III. CROSS-NET DEBUGGER	6
IV. ELF DEVELOPMENT AND PERFORMANCE MEASUREMENT.	8
V. BCPL RUNTIME SUPPORT	9
VI. DIGITAL UNIT ATTACHMENT AND TEST	10
VII. GATEWAY DEVELOPMENT AND DEMONSTRATION.	12
VIII. STATION FORWARDER DEVELOPMENT.	13

I. INTRODUCTION

Progress in Packet Radio areas has covered a broad spectrum at BBN during this quarter. Support materiel has been a target of major effort, both in terms of software development and hardware configuration. Progress in station software includes work on the ELF operating system, which provides the programming environment in which code specific to the packet radio station will execute. During this quarter we entered into extensive cooperative efforts with the developers of ELF to achieve modifications beneficial to both parties. Another item of software progress is the cross-network debugger. This tool provides powerful loading, dumping and debugging capabilities, implemented on a large computer which accesses the smaller PDP-11 packet radio station computer over the ARPA network. Yet another advance in the software environment is the creation of a library package for BCPL, the high level language chosen for implementing programs specific to the packet radio station. Diagnostic and exercise software constitute another area of progress.

Hardware has been configured toward the arrangement desired for a station. The major progress in the hardware area during this quarter is receipt and checkout of a Packet Radio Digital Unit, or PRDU. With this milestone many design issues were confirmed for the first time at BBN. The successful transfer of packets between the station and the PRDU is an encouraging step in station development.

Besides software and hardware support work, this quarter has seen progress on station operation itself. BBN has participated actively in the exchange of information among the Packet Radio Working Group contractors as design issues were resolved. The release of a major document (note 143) specifying basic modules of the station is an important milestone. Two important functions of the station, forwarding packets within the Packet Radio Network, and transmitting packets between the PRN and another network (the gateway function) have been studied in detail. These studies have provided sufficient specificity and insight to permit programming simple, stand-alone versions of these functions. In both cases the functions performed as desired.

II. MEETINGS AND PUBLICATIONS

Two BBN staff members attended a meeting on network security and internetwork end-to-end encryption during the first week of August.

Beginning in May and extending through the remainder of this quarter, a total of 13 weekly progress reports have been written and submitted to the Packet Radio Systems Engineering and Technical Direction (PRSETD) contractor.

Descriptions of the BBN exerciser program for the IMP11-A interface, and of the BBN XNCP (experimental network control program, providing improved access to the ARPANET from programs running under the ELF operating system), were released in June.

Packet Radio Temporary Note 143, "Specification of Basic PRN Station Modules," was released in draft form in June.

SPP (Station to Packet radio unit Protocol) and packet header specifications issued by Network Analysis corporation were reviewed and a response issued in June.

Information on BBN's XNCP was provided to Danny Cohen's group at ISI in June. This was part of an exchange between these two research groups regarding an evaluation of the Virtual Memory ELF operating system.

In July BBN reviewed Collins Radio Corporation's Packet Radio Temporary Note 144. Differences between CRC and BBN packet header

formats were resolved and BBN issued a memo describing the result.

BBN reviewed the ISI memo of June 27, "Comprehensive Measurement Specifications," and responded with an implementation sketch, in July.

Packet Radio Temporary Note 147, documenting BBN modifications to the Virtual Memory ELF operating system, was released in July.

Packet Radio Temporary Note 145, "Cross-Radio Debugging of Packet Radio Units," was issued in July. This document reflected the results of dialogue between BBN and Collins Radio resulting from a release of a preliminary version of this document earlier this quarter.

In July, Packet Radio Temporary Note 148, issued by Network Analysis Corporation, was reviewed. Extensive comments and criticism were prepared and sent to the Packet Radio SETD contractor.

In July, Stanford Research Institute's Packet Radio Temporary Note 149 was reviewed. Comment was found necessary only on some minor issues, and this was communicated in a Packet Radio Weekly Report.

In August, programs to test looping configurations of the PDP-11 and the Packet Radio Unit were delivered to SRI with detailed instructions for their use.

Throughout the quarter, continuing discussions have been held with the Packet Radio SETD contractor on issues of directory functions in the Packet Radio Network, and simulations of the Packet Radio Network. Some of these discussions have been by telephone, while others have been extended text transmitted by network mail.

III. CROSS-NET DEBUGGER

Improvements have been made to X-NET, our cross-network debugger, as their importance was recognized and as time permitted. These include:

- halfword typeout mode
- control-E to abort awaiting network replies
- control-Q to query network reply status
- buffering of type-ahead while awaiting network replies
- four times as many breakpoints (now 32)
- improvement to operation of "end debug" command
- improved printout upon receipt of "asynchronous reply" network messages.

Modification of the BBN XNCP (experimental network control program) to use the IMP11-A interface during this quarter advanced BBN software toward compatibility with the specified Packet Radio Station, which uses an IMP11-A interface instead of the ANTS interface presently on the BBN machine. This also furthered software exchange with SRI, whose machine has IMP11-A interfaces both for ARPANET and Packet Radio traffic.

During June the BBN cross-network debugger was demonstrated to SRI personnel by loading the SRI PDP-11 computer in California with ELF operating system and user program software from BBN in Boston. Operation was error free and response time was good.

Changes to the debug process were made during June to convert

it from operation under non-Virtual Memory ELF to operation under the Virtual Memory ELF operating system. This change allows processes created and debugged by the debugger to no longer potentially conflict with addresses used by the ELF operating system or the debugger itself. In concert with changes to the debug process were changes made to ELF itself. These changes were installation of the "accumulator block" concept whereby separate storage is provided for the values of PDP-11 accumulators during system calls and during execution of user programs. This is necessary so a programmer will obtain a meaningful set of register values at all times.

During the quarter, checkout of the Virtual Memory ELF system and the cross-network debugger progressed. By the end of the quarter, the system and debugger could be used to load and run user programs. BBN interfaced with ELF development personnel elsewhere to iron out problems and forward our modifications to them.

Finally, further improvements were made in the PDP-10 end of the cross-network debugger near the end of this quarter. The debugger now assigns the "special queue" ARPANET host resource on a basis specific to the PDP-11 host which is to be debugged. This permits two programmers to run instances of the debugger simultaneously on the same PDP-10, provided they are debugging separate PDP-11 computers. Other minor improvements were also made.

IV. ELF DEVELOPMENT AND PERFORMANCE MEASUREMENT

We are currently running virtual memory ELF. Several modifications have been made in the kernel both to correct errors in the version originally released by SCRL and to aid in running the cross-net debugger. Both the XNCP and cross-net debugger run in kernel space in VM ELF. All of the functions of the cross-net debugger under non-VM ELF are now available under VM ELF. In addition, the cross net debugger has been modified to allow creation of user address spaces and loading and debugging of programs in these address spaces. To aid in debugging user processes, AC blocks have been added to VM ELF to provide a consistent set of user process registers. When user program execution switches from user mode to kernel mode, the user mode registers are saved for use by the debugger in response to an examine command from the user. We have sent the modifications we made to VM ELF to Dave Retz who has informed us that he will include them in the next ELF release.

V. BCPL RUNTIME SUPPORT

The BCPL runtime library has been coded and checked out. The hand coded portion of the library provides:

- 1) a CREATE routine which creates a new process running a specified BCPL routine in the current user address space
- 2) a routine which accepts vectors of input and output accumulators and an EMT number, and generates the code for the specified ELF system call
- 3) routines for commonly used ELF system calls, which accept the necessary arguments and return values as specified in the ELF system programmers guide.

Another portion of the library was coded in BCPL and provides teletype I/O routines similar to those in TENEX BCPL.

VI. DIGITAL UNIT ATTACHMENT AND TEST

During this quarter we received a Packet Radio Digital Unit (PRDU) from Collins Radio Corporation. The PRDU was mounted in a rack, connected to power, a terminal, and the IMP11-A interface on the station PDP-11. All diagnostics supplied by Collins Radio were run, with various parameter settings. Certain peculiarities arose and were disposed of by consultation with Collins Radio.

Tests using Collins Radio software in the PRDU and BBN software in the PDP-11 were used to test the ability of these two machines to communicate in both directions. Reports of performance were presented to the Packet Radio Working Group through Packet Radio Temporary Notes and discussed with the Packet Radio SETD contractor. Performance was basically satisfactory, with some doubts arising over rate of packet transfer. Further attention to this issue will be given as Packet Radio Station software evolves.

During this quarter, two problems developed with the PRDU which have subsequently been fixed. Each involved the failure of a PRDU printed circuit board; one board was a Programmable Read Only Memory board, while the other was a Random Access Memory module. Repair was effected by returning the faulty boards to Collins Radio.

Near the end of this quarter, the packet looping programs used by BBN to test the Packet Radio Station configuration of PRDU and PDP-11 were released to Stanford Research Institute with detailed instructions for their use. Transfer of this diagnostic and

exercise software furthered the cooperative efforts in software exchange between these contractors. Proper operation of these programs on the SRI hardware, with assistance from BBN as required, will be an important step both in verifying the correct operation of SRI hardware and in demonstrating the compatibility of the SRI and BBN machines.

VII. GATEWAY DEVELOPMENT AND DEMONSTRATION

A simple gateway was coded for the PDP-11. This was checked out with a TCP on one PDP-10 generating messages for another TCP on the same machine and routing these messages through the gateway PDP-11.

VIII. STATION FORWARDER DEVELOPMENT

A simple station forwarder has been coded. This forwarder accepts routing information, packet radio unit identifiers and labels, from the teletype. It receives messages from the packet radio input device (the input side of the IMP11A interface), searches the routing table for the destination PRU id specified in the packet header, inserts the appropriate labels in the packet, and outputs it to the packet radio output device (the output portion of the IMP11A interface). We have checked out the forwarder by modifying it to generate a message and send it to the PRU, which in turn loops it back to the forwarder. We have sent the forwarder to SRI where it will be used in further tests with the appropriate code from Collins.

BBN Report No. 3122

September 1975

COMMAND AND CONTROL RELATED COMPUTER TECHNOLOGY

Part II. Speech Compression

Quarterly Progress Report No. 3

1 June 1975 to 31 August 1975

The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied of the Defense Advanced Research Projects Agency or the United States Government.

This research was supported by the Defense Advanced REsearch Projects AGENCY under ARPA Order No. 2935, Contract No. MDA903-75-C-0180.

Distribution of this document is unlimited. It may be released to the Clearinghouse Department of Commerce for sale to the general public.

I. INTRODUCTION

In our project on digital speech compression, our efforts during the last quarter were primarily concerned with objective speech quality evaluation and real-time implementation of the Linear Predictive speech Compression (LPC) system.

Below we describe a framework that we have chosen to use for objective speech quality evaluation. Within this framework, we developed a number of objective measures which can be computed on-line when the LPC vocoder is in operation. We developed software programs which, in addition to computing these measures, provide a number of statistical data about these measures in the form of line printer outputs and graphical plots. These data have proven to be very instructive in relating the properties of a measure to speech events occurring in an utterance.

We conducted a preliminary study of correlating the objective scores resulting from the use of some of the developed measures with the subjective speech quality evaluation results. As anticipated, any one measure did not always correctly predict the subjective rank ordering of speech utterances produced by different LPC vocoders. Further, even when such correct prediction occurred, the objective scores tended to be very closely clustered indicating the presence of significant "noise" in these scores. Our plan for the next quarter to attempt to resolve this difficult problem is briefly outlined at the end of Section II-C.

As a second concentrated effort in the last quarter, we worked towards bringing up the LPC vocoder on our configuration of PDP-11/SPS-41 system.

II. OBJECTIVE SPEECH QUALITY EVALUATION

The ultimate criterion for determining the quality of the speech that is produced by any compression, encoding or transmission system is the way it sounds to the human listener. Although there are well established procedures to test the intelligibility of speech, little work has been done in developing procedures to test speech quality, and in particular vocoder speech quality. The few procedures that are available are subjective and require extensive testing with human listeners, which is expensive in terms of both time and money.

It would be desirable to develop objective procedures for speech quality evaluation that correlate well with the scores obtained from subjective listening tests. These objective measures would ensure uniformity in evaluation as well as enable the evaluation to be done by computer. Also, the measures can be used in the design of better quality vocoders. While there exist methods in the literature for objectively evaluating the intelligibility of speech in the presence of stationary noise [1,2], little has been done regarding the objective evaluation of either the intelligibility or the quality of vocoded speech. The problem is that if one regards the distortion in the vocoded speech signal as noise superimposed on the signal, then this noise is not only nonstationary but is correlated with the signal. This makes the problem of objective evaluation of vocoded speech quality a difficult one. However, given the immense long-term benefits in

terms of time and money, any headway into the solution of the problem is desirable. For the short term, we hope that objective procedures could at least be used to determine the relative quality of the speech produced by certain vocoder systems.

Before we discuss specific objective measures of speech quality, first we give a detailed exposition of our rationale for the particular course of action we have chosen to develop these measures.

A. A Framework for Objective Speech Quality Evaluation

Any objective measure for the evaluation of vocoded speech quality must be a function of the transmission parameters and must somehow relate to perception. We have based our work in objective quality evaluation on the following observations:

- (1) Speech synthesized from unquantized parameters which are extracted every 10 ms is, for vocoding purposes, practically indistinguishable from the original speech.
- (2) Except for pitch and gain, the fidelity of the spectrum is the principal determiner of quality.
- (3) The spectrum is uniquely defined in terms of the linear prediction filter parameters.

The first observation gives us an anchor point defined in terms of the system parameters and against which to compare quantized realizations of the same utterance. The second and third

observations relate the filter parameters to speech quality through the concept of spectral fidelity. This, then, gives us a framework within which to develop the desired objective measures of speech quality.

Before we proceed there is one important qualification to the preceding discussion. In the first observation above we made the implicit assumption that all the pitch values were correct. While there exist pitch extractors that are very good, we do not know of a "perfect" pitch extractor. In fact, there is not complete agreement on how to exactly evaluate pitch extractors. Because of the elusive nature of this problem, we have decided for the near future to concentrate our efforts on the effects of parameter quantization, time quantization (or frame rate of data transmission) and parameter interpolation on speech quality. Of course, we include pitch as a parameter, but assume that it has been correctly extracted.

One might wonder why we plan to use the unquantized parameters as our reference instead of using the original waveform directly. This question becomes more poignant given the problems with pitch extraction. The answer to this question is obvious once one relates it to speech perception. It is well known that, except for pitch, phase information is quite irrelevant to the perception of speech. This means that two utterances could be perceived the same even if the detailed waveform is quite different. It is difficult to imagine an error criterion on the waveform which would be insensitive to phase. The answer is clearly to go to the spectrum.

In fact, LPC analysis preserves only the magnitude of the spectrum and then uses a minimum phase realization for synthesis. Inasmuch as quality is determined by spectral fidelity, our choice of the unquantized parameters (which determine the spectrum uniquely) as our reference is valid. The implication of such a choice, of course, is that the quality measuring procedures are to be built "inside" the vocoder instead of outside it. Comparisons are made between the unquantized parameters (reference system) and the parameter values used at the synthesizer (test system). We mention in passing that the procedures we develop will be equally applicable to pitch-excited and residual-excited LPC vocoders.

B. Objective Measures of Speech Quality

Given a speech utterance processed by an LPC vocoder, an objective measure summarizes the error or deviation between the reference and the test sets of parameters in terms of a single number which we shall call an objective evaluation score. The objective score would be expected to reflect the perceived quality (relative to the reference) of the speech utterance if, indeed, the objective measures were sensitive to all quality-determining factors. It is unreasonable, and perhaps too simplistic, to expect that one objective measure could always correctly predict perceived speech quality. The chance of such a prediction may be enhanced by combining a number of objective measures in some fashion to obtain an overall objective score. Each measure may be sensitive to some aspects of quality. Ultimately, we plan to perform a

multidimensional analysis on the objective scores obtained from a number of measures with the hope of relating them to different quality dimensions yet to be discovered. For the present study, however, we chose to develop a number of objective measures and investigate each of them separately so as to become familiar with their properties.

For each data frame, an error between the reference and the test parameters can be computed using an appropriate "distance" measure. Ideally, such frame errors should be computed only at selected points in time within the speech utterance that are "perceptually significant". For the purposes of the present study, we simply computed the frame error at a fixed rate, say, every 10 ms. We have thus two problems. (1) To develop suitable distance measures to compute frame errors. This problem is considered below in Section II-B-1. (2) To combine all the frame errors within a speech utterance into one number, which provides the objective score. Different ways of performing this reduction are discussed in Section II-B-2.

1. Objective Error for a Given Frame

We have considered several distance measures for computing the error between the reference and the test parameters of a given frame. In this report, we present a class of spectral measures which are based on the power spectrum of the linear predictor. This is based on our observation that accurate representation of the

short-term speech spectrum is necessary for preserving speech quality. Other measures we are investigating include log area ratio error and log likelihood ratio [3]. We note that many of the considerations discussed in this report for spectral measures apply to these other measures as well.

a. Normalization and Error Definition

The spectrum of the linear predictor is given by [3, 4]

$$P(\omega) = \frac{G^2}{S(\omega)} = \frac{R_o V_p}{\left| 1 + \sum_{k=1}^p a_k e^{-j\omega k} \right|^2}, \quad (1)$$

where G is the linear predictor gain, R_o is the speech signal energy, V_p is the normalized prediction error, $S(\omega)$ is the spectrum of the inverse filter and a_k , $1 \leq k \leq p$, are the predictor coefficients. We will use the subscripts r and t with these quantities to denote the reference and the test cases respectively. Before defining an error function $e(\omega)$ between the power spectra $P_r(\omega)$ and $P_t(\omega)$, it may be desirable to normalize these spectra in some fashion. For instance, they may be normalized to have the same total energy, R_o . Since the total energy is the same as the arithmetic mean of the spectrum, i.e.,

$$R_o = \frac{1}{2\pi} \int_{-\pi}^{\pi} P(\omega) d\omega, \quad (2)$$

we shall call this procedure as arithmetic mean (AM) normalization. A second type of normalization, called geometric mean (GM) normalization, equates the geometric means of the two spectra. The GM of the power spectrum in (1) can be shown [4] to be equal to $V_p R_o$, i.e.,

$$GM = \exp \left[\frac{1}{2\pi} \int_{-\pi}^{\pi} \log P(\omega) d\omega \right] = V_p R_o \quad . \quad (3)$$

Clearly, AM normalization results in the spectra having the same area in the (linear) spectral domain, while GM normalization causes the areas under the spectra in the log spectral domain to be equal. While these two are the types of normalization most commonly used, certainly other types of spectral means can be used for normalization. (For a definition of these means, see the next subsection.)

The error function between the normalized spectra can be defined either in the spectral domain as

$$e(\omega) = P'_r(\omega) - P'_t(\omega) \quad , \quad (4)$$

or, in the log spectral domain as

$$e(\omega) = \log P'_r(\omega) - \log P'_t(\omega) \quad , \quad (5)$$

where the primes are used to indicate normalized quantities. Other reasonable error definitions include

$$e(\omega) = \frac{P'_r(\omega) - P'_t(\omega)}{P'_r(\omega)} \quad , \quad (6)$$

$$e(\omega) = P'_r(\omega)/P'_t(\omega) \quad . \quad (7)$$

AM normalization combined with the error definition (4) gives

$$e(\omega) = \left[V_{pr}/S_r(\omega) \right] - \left[V_{pt}/S_t(\omega) \right] \quad . \quad (8)$$

Using GM normalization in (5) we get

$$e(\omega) = \log[S_t(\omega)/S_r(\omega)] \quad . \quad (9)$$

A "hybrid" combination which we will have occasion to refer to later on is obtained with AM normalization and the error definition (5):

$$e(\omega) = \log(V_{pr}/V_{pt}) + \log[S_t(\omega)/S_r(\omega)] \quad . \quad (10)$$

Other error expressions can be similarly derived using different combinations of spectral normalization and error definition.

The various choices of normalization and error definition in forming an objective measure may relate differently to quality-determining factors. We are currently in the process of

comparing the properties of some of these choices (see Section II-C).

b. Weighted Error Norm

Given the error $e(\omega)$ as a function of ω , we wish to compute the norm of the error. We note that $e(\omega)$ is computed in practice at a set of, say, N discrete frequencies. A general weighted k -th norm, known as the L_k norm, is given by

$$L_k = \left[\frac{\sum_{i=1}^N w(i) |e(\omega_i)|^k}{\sum_{i=1}^N w(i)} \right]^{\frac{1}{k}}, \quad w(i) > 0, \quad 1 \leq i \leq N, \quad (11)$$

where $\{w(i)\}$ is the sequence used to weight the individual errors. L_k is sometimes called the weighted generalized mean of order k . For the moment, let us assume that the weights are all equal to 1, i.e., $w(i) \equiv 1$. (Different weighting methods are discussed later on in this subsection.) Then, we note that L_{-1} is the harmonic mean, L_0 is the GM, L_1 is the AM, and L_2 is the root mean square value of the sequence of absolute error values. Another property of the above error norm is that

$$\begin{aligned} L_\infty &= \max_i |e(\omega_i)|, \\ L_{-\infty} &= \min_i |e(\omega_i)|. \end{aligned} \quad (12)$$

Between this minimum and maximum values, L_k is a monotonically

increasing continuous function of k .

These properties may be interpreted in the following manner with respect to averaging of errors. As the order of the norm is reduced, the resulting average becomes more representative of the smaller errors in the sequence considered. Conversely, as the order of the norm is increased, larger errors are being more strongly emphasized. One may argue that large errors are the primary influence in speech quality decisions, therefore justifying a large order for the error norm. Another viewpoint may be that a few spurious large errors may not affect the subjective quality judgments as much as the level and amount of smaller amplitude errors. The choice of the order of the error norm has not yet been resolved. We are continuing to investigate the significance of different ordered norms.

In forming the error norm given by (11), some errors may be emphasized more than others by differentially weighting the individual errors. If the weighting sequence $\{w(i)\}$ is determined based on some concept of speech perception, the resulting error norm should presumably enhance sensitivity to quality-determining factors. Below we describe two specific weighting methods that we have used. It should be clear that several composite weighting functions may be obtained from these two weighting functions. For example, $w(i) = w_1(i)w_2(i)$, where w_1 and w_2 are two weighting functions, is a new weighting function having its own properties different in general from those of w_1 and w_2 . Since each weighting

function may emphasize a certain aspect or set of aspects considered important for perceived quality, one hopes to obtain a weighted error norm that is sensitive to most of the quality-determining factors by combining two or more of these weighting functions as mentioned above.

(i) Spectral Intensity Weighting

Since high spectral intensities generally occur close to formant peaks, spectral errors at the corresponding frequencies may be of greater perceptual significance and hence in forming the error norm it is reasonable to emphasize these errors more than those that occur at low spectral intensities. (We note that the error definition given by (8) already incorporates a similar relative emphasis.) We have considered a weighting function of this sort whose value at frequency ω_i is given by the ratio of the spectral intensity at this frequency to the peak intensity in the spectrum:

$$w(i) = P_r(\omega_i)/P_{\text{peak}} \quad (13)$$

We intentionally used the reference spectrum $P_r(\omega_i)$ in (13) in defining the weighting function to ensure identical weighting when comparing different LPC vocoders using the same speech utterance processed by them. Although the weighting factor in (13) gives the intuitively appealing relative power level, it is clear from (11) that exactly the same effect is achieved by weighting directly with $P_r(\omega_i)$. We use the latter method since it avoids the computation of

P_{peak}

(ii) Articulation-Index Based Weighting

French and Steinberg [5] have developed a physical measure called the Articulation Index, or AI, which is highly correlated with speech intelligibility as evaluated by subjective speech perception tests. Adapting some of the results used in AI computation, we have derived a weighting function which decreases exponentially with frequency. Since it is not unrealistic to consider speech intelligibility and quality as related phenomena, use of this function for weighting the spectral errors may be expected to enhance the sensitivity of the error norm to perceived quality.

AI is a weighted fraction representing, for a given speech channel and noise condition, the effective proportion of the normal speech signal which is available to a listener for conveying speech intelligibility. AI is computed from acoustical measurements or estimates, at the ear of a listener, of the speech spectrum and of the effective masking spectrum of any noise which may be present. An important point to note is that speech intelligibility can be validly evaluated by the AI procedure only for speech corrupted by stationary noise. Such an evaluation cannot be performed, for example, for vocoded speech where the superimposed noise is nonstationary.

Since the details of AI computation are quite involved, we present below only the concepts that are directly needed in our derivation. For computational convenience, one divides the spectrum of speech and noise into a finite number of frequency bands. The AI concept holds that any one frequency band carries a contribution to the total AI which is independent of the other bands and that the total AI is given by the sum of the contributions of the separate bands. If ΔA_i represents the contribution of the i -th frequency band and N the number of bands, then

$$AI = \sum_{i=1}^N \Delta A_i, \quad 0 \leq AI \leq 1 \quad (14)$$

ΔA_i takes values from zero to a maximum of $(\Delta A_i)_{\max}$ as the absolute levels of speech and noise at the listener's ear are independently varied over wide ranges. If W_i represents the fractional part of $(\Delta A_i)_{\max}$ which is contributed by the i -th band with a particular combination of speech and noise, then $\Delta A_i = W_i (\Delta A_i)_{\max}$. Detailed procedures are available to compute $\{W_i\}$ for any given speech channel and noise condition [5].

If conditions are optimum for hearing speech in all the bands, then $W_i = 1$ for all i . For this case, the AI of low-pass filtered speech of an adult male speaker has been experimentally determined as a function of the cut-off frequency of the low-pass filter as shown by the solid curve in Fig. 1 [6]. Notice that speech low-pass filtered at or below 200 Hz has zero AI (i.e., no intelligibility),

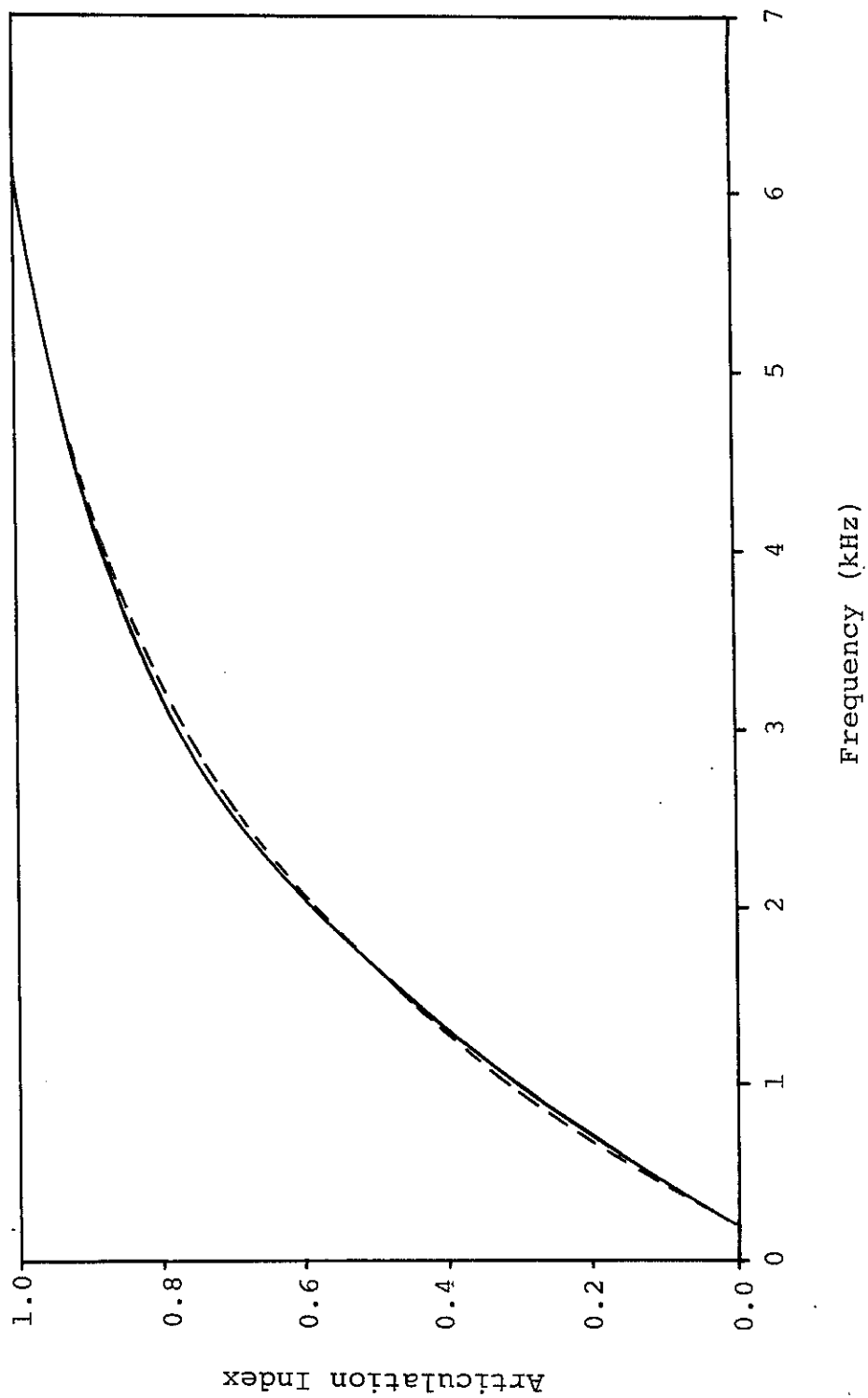


Fig. 1. Solid Curve: Articulation Index versus frequency when all bands are at their optimum levels.
Dashed Curve: Analytic approximation to the solid curve.

while speech low-pass filtered at or above 6100 Hz has an AI of unity (i.e., complete intelligibility). The derivative or slope of this function at any frequency shows the importance of that frequency with respect to its maximum possible contribution to AI. At any frequency the product of the slope of this curve and the factor W at this frequency represents the actual contribution of this frequency to the total AI. An inspection of the solid curve in Fig. 1 indicates that lower frequency spectral values have a potential to contribute more to AI and hence to intelligibility than higher frequency spectral values.

From the above considerations, we chose to use the derivative of the solid curve in Fig. 1 for weighting the spectral errors. Instead of computing the derivative directly from the AI curve we decided to fit it with an analytical function and then use its derivative in the computation of a weighting function. A reasonable fit to this curve was obtained with the following exponential function:

$$\hat{A}(\omega) = \begin{cases} \frac{1 - e^{-\alpha(\omega - \omega_1)}}{1 - e^{-\alpha(\omega_2 - \omega_1)}} & , \omega \geq \omega_1 \\ 0 & , \omega < \omega_1 \end{cases} \quad (15)$$

where $\omega_1 = 200$ Hz and $\omega_2 = 6100$ Hz are the lower and upper frequency limits of the AI curve in Fig. 1. α was determined by a least squares fit of $\hat{A}(\omega)$ to data from the solid curve of Fig. 1. The optimal value of α was found to be 68×10^{-6} . The resulting $\hat{A}(\omega)$ is shown in Fig. 1 as the dashed curve. We define the AI-based

weighting function as being proportional to the derivative $d\hat{A}(\omega)/d\omega$:

$$w(\omega) = \beta e^{-\alpha\omega}, \quad \alpha = 68 \times 10^{-6} \quad (16)$$

The actual value of β in (16) is not important as it gets cancelled when we substitute (16) into the error norm expression (11). Fig. 2 depicts the AI weighting function with β chosen such that its value at 5 kHz is unity. We have plotted the weighting function up to 5 kHz only since our spectral analysis is limited to this frequency. Although speech spectrum below 200 Hz does not contribute to AI (and hence to intelligibility) as shown in Fig. 1, it may however affect speech quality. A convenient way of obtaining a weighting function for this spectral region is to simply extend the definition (16) below 200 Hz as shown by the dashed line in Fig. 2. The AI weighting function thus decreases exponentially with frequency. When it is used in the error norm (11), low frequency spectral errors will be emphasized more than the high frequency spectral errors.

2. Objective Error for a Speech Utterance

As mentioned earlier, we compute the frame spectral error every 10 ms, thus generating a frame error sequence $\{E(j)\}$. If the LPC vocoder under test had transmitted the parameter data for the j -th frame, then the frame error $E(j)$ would be due to parameter quantization only. In the case when there was no data transmission for the j -th frame, then $E(j)$ would contain errors due to both

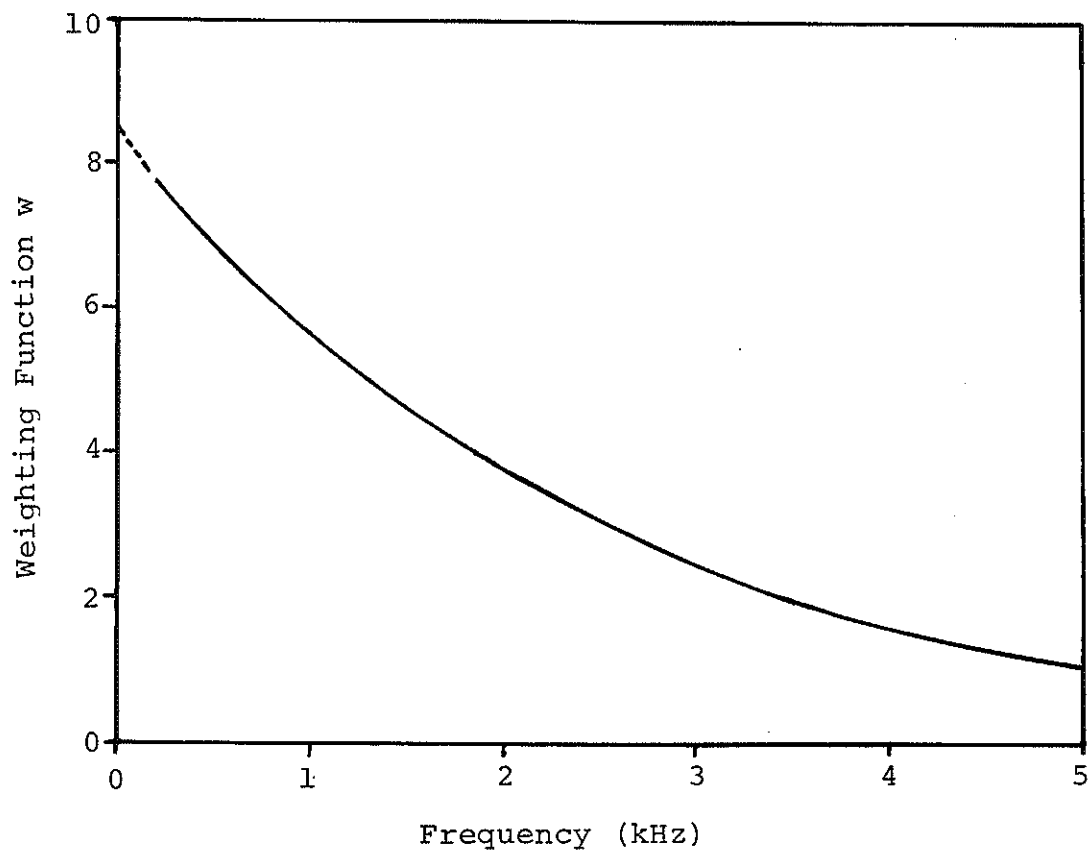


Fig. 2. AI-based weighting for the spectral error. The weighting function shown was normalized such that its value at 5 kHz is unity.

parameter quantization and interpolation. (We have assumed the transmission channel to be error-free.)

Given the error sequence $\{E(j)\}$ for a speech utterance, the problem is to compute an objective score that summarizes these errors in some fashion. Weighted averaging as discussed in Section II-B-1b generates a class of procedures to do this task. We discuss below two weighting functions that we have considered. Alternatively, many other statistical parameters such as median, mode, etc., of the weighted frame errors can be used to obtain objective measures. This topic is considered later on.

a. Weighted Averaging of Frame Errors

The weighted average of the frame error sequence is computed using the previously defined error norm (11) where $\{e(\omega_i)\}$ is now replaced with $\{E(i)\}$. The weighting sequence $\{w(i)\}$ in (11) employed for this computation has a different meaning: It corresponds to error weighting in time. Such weighting is based on the assumption that errors in certain segments of a speech utterance are relatively more important to perceived speech quality than those in other segments. We have considered in our investigations two specific weighting functions described below. (As before, composite weighting functions can be obtained from these two.)

(i) Filter Gain Weighting

It is reasonable to assume that frame errors in low energy regions of an utterance do not contribute as much to quality judgment as those in high energy regions. For example, large changes in the spectrum may not often be detected by the listener if the total energy in the spectrum is low. Based on this, one can consider a simple scheme which weights the frame error proportional to the corresponding filter gain in decibels, i.e.,

$$w(j) = c \log G_j, \quad (17)$$

where the constant c , for instance, may be chosen such that $0 \leq w(j) \leq 1$.

An example of how this weighting scheme reduces errors in low energy regions is illustrated in Fig. 3. The plot of filter gain versus time is given in Fig. 3a. The solid curve in Fig. 3b is the unweighted frame error as a function of time, and the dashed curve shows the gain-weighted error. The effect of the gain weighting is especially apparent in the low energy region 0.1-0.2 sec.

While the simple gain weighting scheme in (17) is effective in essentially ignoring errors that occur during silences, it is not adequate in general. For instance, it overly deemphasizes the frame errors corresponding to low energy regions even though some of these regions (e.g., stop bursts) may contain vital cues to speech perception. A better weighting scheme may be obtained by making the

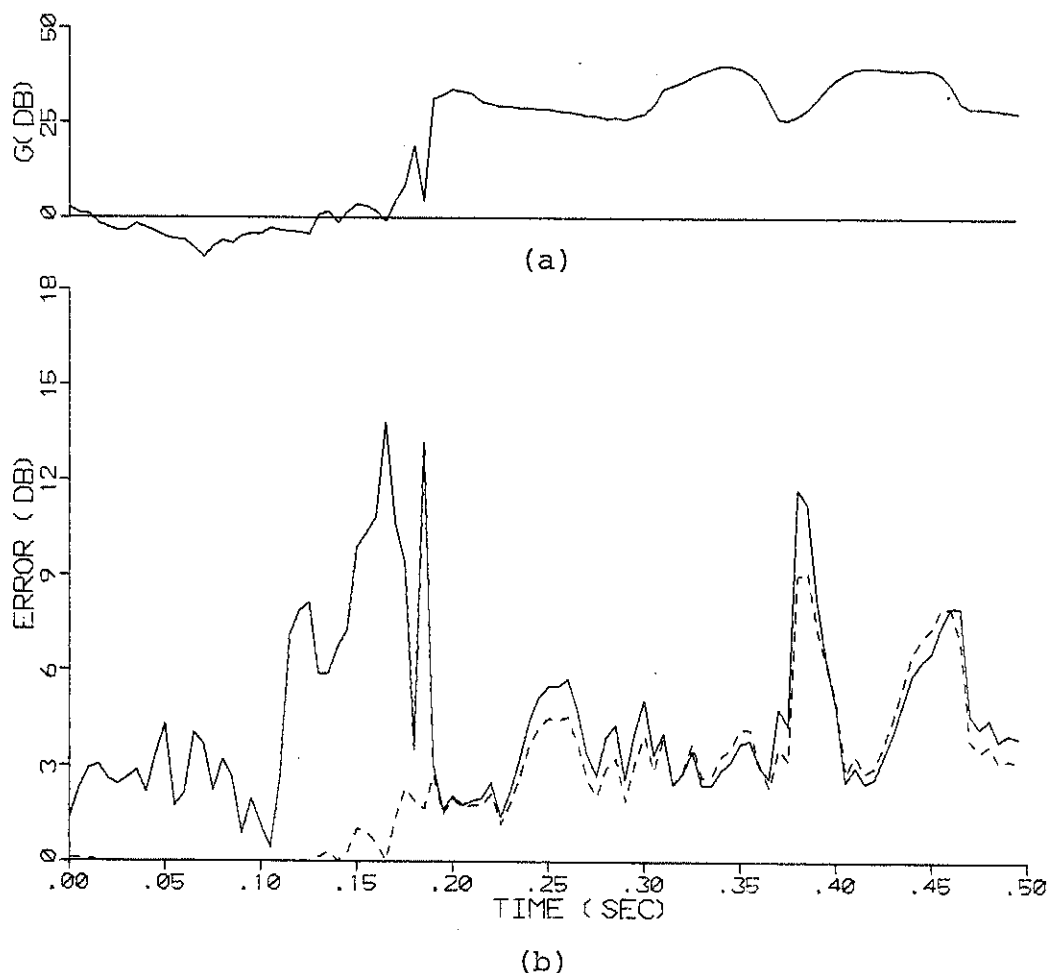


Fig. 3. (a) Filter gain versus time. (b) Frame error versus time. Solid Curve: The unweighted L_2 -norm of the error defined in (10). Dashed Curve: Same as solid curve except that filter gain weighting in time has also been applied. (We have used $10 \log_{10}$ instead of the natural logarithm in evaluating (10), which explains why the error-axis in Fig. 3b is labelled in decibels.)

weighting factor depend nonlinearly or piece-wise linearly on the filter gain in decibels. The specific functional relationship should be chosen so as not to destroy any quality-bearing information present in the frame errors. This problem is currently under investigation.

(ii) Spectral Difference Weighting

Another consideration relevant to speech quality is the rate at which speech characteristics change in time. Errors in the synthesized speech during steady state sounds may not be as important to the perceived quality as those which occur during rapid speech changes such as in phoneme transitions. In order to account for this effect, we plan to determine the rate of change of the spectrum in terms of a suitable spectral difference measure, and use it for weighting the frame error appropriately. We are currently investigating a number of ways of defining a spectral difference measure.

b. Statistical Measures of Frame Errors

Although an averaged frame error is important, this measure by itself may not be sufficient to describe speech quality. For example, it may be important to know what amplitude of error occurred most often and, whether or not this error coincides with the average. Studies of the statistical properties of the frame error of this sort may prove instructive in developing new measures and improving the average error measure. We, therefore, have begun

observing characteristics of error distributions.

Software has been developed to compute the median, mode, deciles and other characteristics of the distribution of the frame error. We are also observing histograms of the frame errors in the hope that their shapes and characteristics will provide insight needed to understand quality-determining factors. For instance, from the error distributions one may be able to define thresholds which could be used to prevent small errors from entering into the average. An example where this approach may be appropriate is illustrated in Fig. 4. The error distribution in Fig. 4a was computed without using gain weighting, while gain weighting was applied for the error distribution of Fig. 4b. A large concentration of errors near zero is evident for the gain-weighted error. One may decide that these small errors, say below about 3 dB or so are not important and could be discarded. The resulting errors could then be used to compute a new distribution with a new median, mode, average, etc. These new quantities may be more significant since they represent errors during relatively loud segments in the utterance. This approach will be studied further to determine its merit in quality evaluation.

Although this portion of our study is in its preliminary stages, it is apparent that error distributions are useful in understanding the factors influencing subjective speech quality.

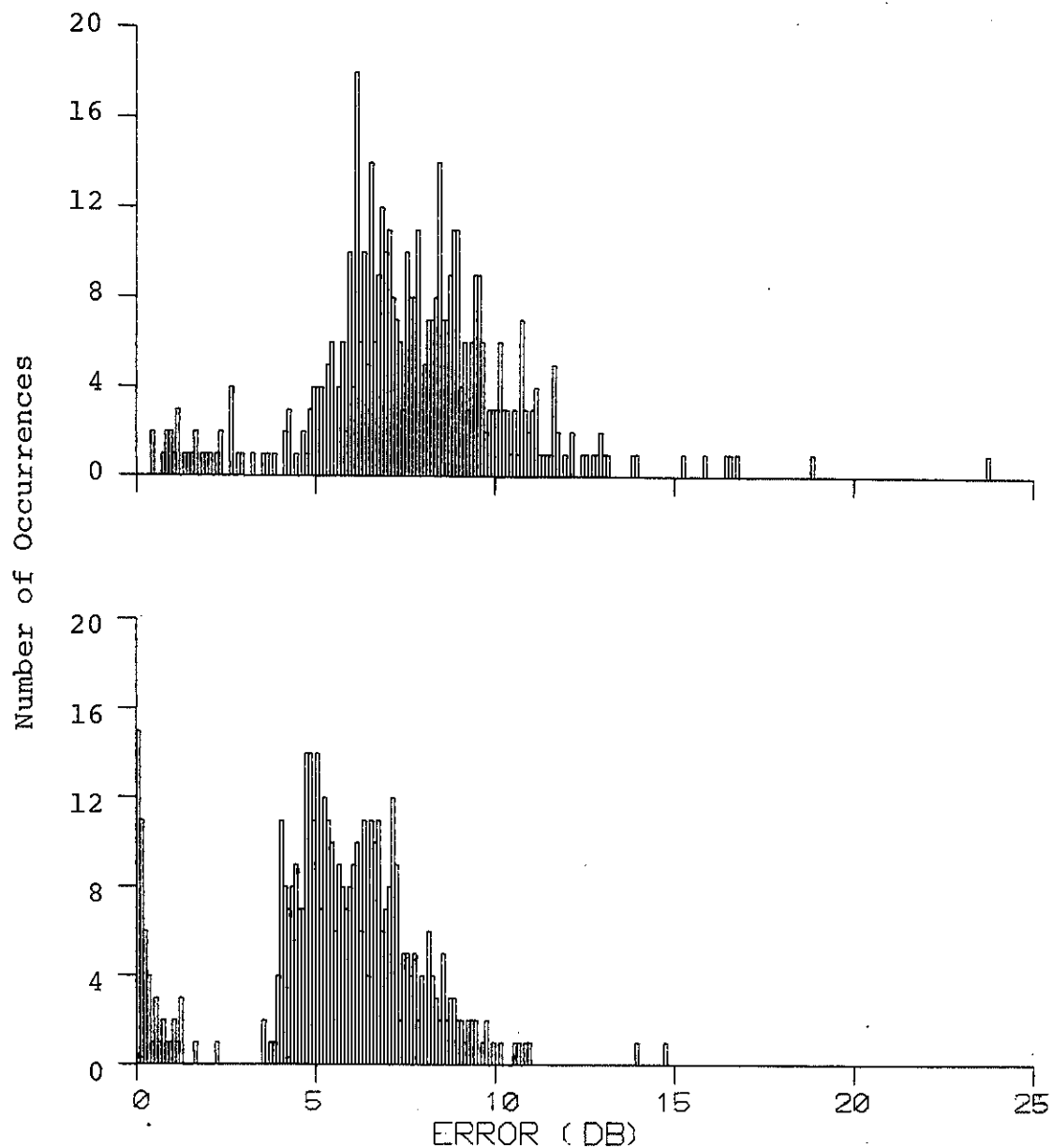


Fig. 4. Histograms of frame errors. (a) Frame error was computed using the L₂ norm of the error defined in (10). The product of the spectral intensity and AI weights was used for weighting the spectral errors. (b) Same as (a) except that filter gain weighting in time has also been applied.

C. Preliminary Testing of Some Objective Measures and Future Plan

All the objective measures we discussed above have been implemented and their properties are being studied. We made some preliminary tests correlating the objective scores obtained using some of the developed objective measures with the subjective quality evaluation results. (It should be pointed out that in this preliminary study we did not consider errors in pitch and filter gain.)

We considered the two utterances JB5 (male: "The little blanket lay around on the floor.") and RS1 (female: "Why were you away a year, Roy?") processed through each of the two LPC vocoders #6 and #12, which had been tested in our subjective speech quality evaluation project [7]. Thus, the subjective speech quality scores for these four vocoded utterances were already available. We computed the objective scores using a number of objective measures discussed in the previous sections. Comparisons of the objective and the subjective scores indicated that no single objective measure was able to always correctly predict the subjective rank ordering of vocoded speech utterances. Further, even when such correct prediction occurred, the objective scores tended to be very closely clustered indicating the presence of significant "noise" in these scores. We illustrate this latter point by presenting the results of a test in which we used the following measure: The frame spectral

error was determined as the L_1 -norm of the unweighted error defined in the log domain after AM normalization (see (10)); time averaging of the unweighted frame errors was performed using L_2 -norm. Table 1 lists both subjective and objective scores for the four utterances. This particular test indicates that the objective measure used is able to correctly predict the subjective rank ordering of the vocoded speech utterances considered. However, while the subjective quality ratings for the test utterances given in Table 1 were relatively widely separated (see Fig. 4, [7]), the measured objective scores (average errors) varied by less than 10%. In conclusion, the tested objective measures are not sufficiently sensitive to perceived quality-determining factors. In retrospect, it was probably unrealistic to expect good results from such simple measures.

In developing better objective measures, not only must we identify the positive factors influencing speech quality but also the noise-contributing elements must be found and eliminated to assure valid and sensitive measures. To this end, we plan to begin a step-by-step program to discover the quality-determining factors in the spectrum for each point in time. Following that, we shall attack the more difficult problem of discovering the time-dependent factors which determine quality.

TABLE 1

Sentence #	Vocoder #	Subjective Quality Scores	Objective Score (Average Error in dB)
RS1	6	-.3 (Good)	3.91
RS1	12	.18 (Poor)	4.16
JB5	12	-.15 (Good)	3.80
JB5	6	.53 (Poor)	4.18

III. REAL-TIME IMPLEMENTATION

During the past quarter, we expended a considerable and concentrated effort in trying to bring up the real-time LPC vocoder on our SPS-41 machine.

We modified the LPC programs and support software supplied by other ARPA-sponsored sites and by SPS, Inc., to run on our configuration of the PDP-11/SPS-41 system. We developed a procedure for loading these programs from TENEX into the PDP-11. We worked towards locating and describing hardware problems in the SPS-41, which appear to be the cause of system failures after short periods of successful operation.

As part of this effort, our SPS-41 machine was moved back to SPS, where we have one person working full time trying to resolve these problems with the help of people from SPS. People from other ARPA-sponsored sites have been cooperating through telephone consultations on this on-going endeavor.

REFERENCES

1. Beranek, L. L., Acoustics, New York: McGraw-Hill Book Co., 1954.
2. Kryter, K. D., The Effects of Noise on Man, New York: Academic Press, 1970.
3. Makhoul, J., R. Viswanathan, L. Cosell and W. Russell, National Communications with Computers, Vol. II, Speech Compression Research at BBN, Report No. 2976, Bolt Beranek and Newman Inc., Cambridge, Massachusetts, December 1974.
4. Makhoul, J., "Linear Prediction: A Tutorial Review", Proc. IEEE, Vol. 63, pp. 561-580, April 1975.
5. French, N. R. and J. C. Steinberg, "Factors Governing the Intelligibility of Speech Sounds", J. Acoust. Soc. Amer., 1947, Vol. 19, No. 1, pp. 90-119.
6. Beranek, L. L., "The Design of Speech Communication Systems", Proc. IRE, Vol. 35, September 1947, pp. 880-890.
7. BBN Quarterly Progress Report on Command and Control Related Technology, Part III, Report No. 3093, June 1975.

BBN Report No. 3122

September 1975

COMMAND AND CONTROL RELATED COMPUTER TECHNOLOGY

Part III. Vocoder-Speech Evaluation

Quarterly Progress Report No. 3

1 June 1975 to 31 August 1975

The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied of the Defense Advanced Research Projects Agency or the United States Government.

This research was supported by the Defense Advanced Research Projects Agency under ARPA Order No. 2935. Contract No. MDA903-75-C-0180.

Distribution of this document is unlimited. It may be released to the Clearinghouse Department of Commerce for sale to the general public.

TABLE OF CONTENTS

	Page
III. VOCODED-SPEECH QUALITY EVALUATION	1
1. INTRODUCTION.	1
2. RATING PROCEDURE.	1
2.1 Rationale.	2
2.2 Procedure for Rating Task.	5
a. Preparation of the Stimulus Tape.	5
b. Collection of Data.	6
c. Analysis of the Rating Data	9
d. Reliability of the Data	10
3. MULTIDIMENSIONAL SCALING PROGRAMS	18
4. PHONEME SPECIFIC TESTS.	18
5. PAPERS.	19
6. FUTURE PLANS.	20
REFERENCES.	22
APPENDIX A - INSTRUCTIONS FOR QUALITY DEGRADATION RATINGS	
	A-1

III. VOCODED-SPEECH QUALITY EVALUATION

1. INTRODUCTION

During the quarter, work on quality evaluation has progressed along several lines. Another subjective evaluation procedure (a rating procedure) has been tested with promising results. The multidimensional scaling programs have been further modified and made more useful. Some progress has been made on the phoneme specific tests, although the tests themselves have been deferred until the next quarter. A paper reviewing comprehensively quality evaluation techniques is in preparation. Another paper describing some of our work to date has been prepared for presentation at the upcoming meeting of the Acoustical Society of America. Each of these activities is discussed in what follows.

2. RATING PROCEDURE

There are two main aims of our work in quality evaluation of vocoder systems. The first is to try to establish the psychological dimensions underlying the subjective evaluation task, and develop testing procedures that make optimal use of this structure. The second aim is to apply these procedures to the LPC-vocoder systems developed by the various contractors for use on the ARPANET.

We described in our first QPR the six sentences, the six talkers, and the fourteen vocoder systems we selected for use in

studying the task of subjective quality evaluation. We also described two rank-ordering tasks, one an automated procedure for ranking triplets of the sentences, and a second in which subjects ranked all fourteen systems for each of the 36 sentence-speaker combinations. Our initial analyses of the data collected with the latter procedure were described in the second QPR. In the present report, we describe a third subjective quality-evaluation task, in which subjects rated the quality of each of the fourteen vocoder systems, for each of the 36 sentence-speaker combinations.

2.1 Rationale

The subjective task that imposes fewest constraints on a listener evaluating speech quality is the paired comparison task. Pairs of stimuli are presented, and the listener simply indicates which member of each pair he prefers. The members of successive pairs may differ in different ways, but since only two stimuli are presented at a time, the listener never has to face this problem. Unfortunately, the number of paired comparisons that has to be made increases with the square of the number of stimuli, so that the procedure becomes unmanageable when there are more than 15 to 20 stimuli to be compared. Since we had generated 504 stimuli, an exhaustive paired comparison test would have required some 250,000 paired comparisons or 500,000 stimulus presentations, corresponding to about six months of full-time effort for each subject, making 250 comparisons per hour! Our first rank-ordering task, with triplets, was designed to reduce the problem of collecting the data to

manageable proportions: the initial test of this task we ran (QPR 1) did not adequately compare each system with each other system. Conversely, an adequate set of cross-comparisons would have failed to reduce the number of judgments required to a manageable size.

Our next task was a ranking task in which subjects were given all 14 versions of a given speaker-sentence combination, and had to rank the 14 systems that produced them. Here, the listener is forced to ignore the fact that different systems can sound different in different ways, and must effectively line the stimuli up on a single quality dimension. Thus we reduce the amount of data to be collected, but at the expense of making the listeners task considerably more difficult, and introducing the risk of reduced reliability. At the same time, the ranking task retains some of the desirable features of the paired comparison task. The listener is free to hear each stimulus or pair of stimuli as often as he likes, and he tends to build up his rank order, by starting with a pair, then placing the third stimulus in the correct ranking with respect to the preceding two, and so on. Thus new stimuli are added by a series of paired comparisons with the members of the existing rank order. This procedure reduces the number of stimulus presentations to perhaps five or ten per stimulus, or a total of 5000 for our stimulus set. Our subjects took between 6 and 10 hours each to rank order the 14 systems for each of the 36 speaker-sentence combinations. A serious disadvantage of the ranking task results from the fact that the range of qualities encountered within the 14 versions of one speaker-sentence combination may be very different

from the range of a second combination. Thus there may be a considerable range of qualities associated with rank 14, but there is no way for the subject to express this although he might be willing and able to do so if given the chance.

The rating task avoids these problems, and is probably the most efficient task possible, in that it requires only a single presentation of each of the 504 stimuli. But at the same time it requires most from the subject. The subject assigns numbers that reflect his perception of speech quality, and he must stick to the same rating system through the whole experiment. If his criterion drifts, as it usually will during an experiment lasting an hour or more, there is no direct way of assessing how much drift has occurred, or of correcting for it. The procedure effectively forces the subject to place all 504 stimuli on a single undimensional scale. Since the stimuli fall naturally into a 6x6x14 matrix (i.e., speakers x sentences x systems), multidimensional scaling methods can be used to extract any structure that may be present in the data. Furthermore, if the structure extracted from the rating data agrees reasonably with that extracted from the ranking data, this supports the idea that the two tasks are being performed on the basis of a single underlying psychological structure. If this can be confirmed, separate rating tasks might be designed to place the stimuli on each of the perceptual dimensions that define the psychological structure. The possibility of establishing the perceptual dimensions and of accurately placing the vocoder systems along them, significantly raises the chances that objective tests

could be developed to assess speech quality, and thus eliminate the need for further subjective testing.

2.2 Procedure for Rating Task

a. Preparation of the Stimulus Tape

It is known that judgments made on speech materials exhibit strong sequential effects (Huggins, 1968). Therefore, it was felt to be important to control the presentation sequence in such a way that the 503 sequential pairs of stimuli occurring in the sequence of 504 items, should be approximately balanced for all possible pairs of speakers, for all possible pairs of sentences, and for all possible pairs of systems. Exact balancing was not possible, since this would have required a multiple of 30 stimulus presentations for speakers and for sentences (with the constraint that no speaker or sentence follow itself), and a multiple of 196 presentations for systems. A presentation order was derived from the experimental design for the ranking experiment (QPR 1, Part III, Table 5, p35), by retaining the six-item sequences specified by the columns in each of the six 6x6 matrixes, and arranging the order of these six-item sequences to minimize imbalance. At the same time, the order was approximately balanced for serial order. That is, each speaker and sentence occurred early in the sequence about the same number of times that they occurred late. The 14 systems were then assigned by a computer program that aimed at an approximate counterbalancing of sequence effects for the systems. The number of times that one system followed a second system ranged from two to four over the 504

item sequence. The complete sequence is given in Table 1.

Generating the stimulus tape from the digital waveform files stored in the computer seemed to have minor advantages over generating the tape by hand directly from the Language-Master cards used in the Ranking experiments, but would have been considerably more expensive. Therefore, we used the latter procedure, with each stimulus presentation beginning 5 1/2 seconds after the start of the preceding stimulus. The sentences were recorded in blocks of 12, with a 10 second break between blocks. The 504 item sequence lasted about 50 minutes.

b. Collection of Data

The same four subjects served for the rating task as had already served in the ranking task. There were two advantages of this: first, all the subjects were already very familiar with the stimulus materials, which minimized any start-up or familiarization effects that might otherwise have occurred. Secondly, using the same subjects permits each subject to act as his own control, and the results of the two tasks (ranking and rating) can be compared for each subject.

The stimuli were presented at a comfortable listening level (about 70 dB SPL) through a high-quality loudspeaker in a sound treated room. The four subjects served simultaneously, and wrote their ratings in a booklet with a new page for each block of 12 stimuli.

Table 1

Stimulus Order for 504 Item Rating Tape
 Spkr(A,B,K,D,S,F), Sent(1-6), System(1-14). One Row per Block
 A4-01 Means Talker A, Sentence 4, System 1.

A4-01	B3-14	K5-14	S2-01	D6-01	F1-02	B6-01	S5-13	F1-14	A4-02	K2-14	D3-03
K4-01	F3-12	B2-14	D5-04	A6-14	S1-05	D3-01	K5-03	A1-14	F4-06	S2-02	B6-02
A2-03	R1-02	K3-13	S6-01	D4-04	F5-01	S6-11	A2-01	D4-05	R1-03	F5-03	K3-04
B2-02	S1-04	F3-13	A6-13	K4-02	D5-05	S3-14	A5-13	D1-12	B4-01	F2-10	K6-01
F2-06	D1-01	S6-07	K3-01	B4-09	A5-01	B4-08	S3-13	F5-04	A2-04	K6-12	D1-13
F6-03	D5-13	S4-05	K1-02	B2-06	A3-14	K5-12	F4-02	B3-07	D6-14	A1-11	S2-14
F3-10	D2-02	S1-12	K4-12	B5-03	A6-05	D2-13	K4-06	A6-03	F3-11	S1-13	B5-07
S1-11	A3-12	D5-06	B2-13	F6-11	K4-03	F5-06	D4-12	S3-04	K6-11	R1-04	A2-10
S2-13	A4-10	D6-12	B3-11	F1-11	K5-10	B1-14	S6-08	F2-14	A5-07	K3-12	D4-07
S4-13	A6-08	D2-12	B5-05	F3-09	K1-03	A5-12	B4-10	K6-10	S3-11	D1-09	F2-02
A3-08	B2-01	K4-04	S1-09	D5-14	F6-09	K3-11	F2-05	B1-05	D4-06	A5-11	S6-06
B5-10	S4-03	F6-10	A3-04	K1-05	D2-11	A6-02	B5-09	K1-13	S4-09	D2-04	F3-08
D5-11	K1-07	A3-02	F6-14	S4-01	B2-03	F1-07	D6-10	S5-05	K2-04	B3-03	A4-08
B3-10	S2-06	F4-04	A1-06	K5-09	D6-05	K1-08	F6-08	B5-06	D2-08	A3-09	S4-12
D6-09	K2-06	A4-05	F1-10	S5-09	B3-09	K2-10	F1-08	B6-07	D3-09	A4-07	S5-03
K6-09	F5-08	B4-05	D1-07	A2-08	S3-03	F4-01	D3-02	S2-11	K5-08	B6-04	A1-07
D1-06	K3-06	A5-14	F2-13	S6-02	B4-02	A1-10	B6-03	K2-13	S5-01	D3-05	F4-12
S5-08	A1-02	D3-13	H6-05	F4-14	K2-02	D4-01	K6-14	A2-05	F5-13	S3-06	R1-07
K6-07	F5-05	B4-03	D1-14	A2-14	S3-07	A1-04	B6-06	K2-12	S5-02	D3-04	F4-03
B1-12	S6-14	F2-03	A5-10	K3-07	D4-14	F3-04	D2-01	S1-10	K4-05	B5-12	A6-01
K3-09	F2-12	B1-06	D4-02	A5-09	S6-13	B4-14	S3-12	F5-07	A2-13	K6-03	D1-02
D5-03	K1-04	A3-13	F6-04	S4-14	B2-08	S5-14	A1-09	D3-14	B6-10	F4-07	K2-11
A4-14	B3-06	K5-01	S2-12	D6-13	F1-09	D2-10	K4-08	A6-12	F3-03	S1-08	B5-01
F6-01	D5-07	S4-10	K1-14	B2-11	A3-01	S2-08	A4-03	D6-11	B3-02	F1-05	K5-02
A4-06	B3-13	K5-13	S2-10	D6-02	F1-12	B6-12	S5-04	F1-04	A4-12	K2-09	D3-06
K4-07	F3-07	B2-12	D5-08	A6-04	S1-07	D3-08	K5-11	A1-03	F4-05	S2-09	B6-08
A2-02	B1-11	K3-10	S6-04	D4-11	F5-09	S6-03	A2-06	D4-10	R1-13	F5-11	K3-08
B2-10	S1-01	F3-06	A6-06	K4-11	D5-12	S3-10	A5-06	D1-08	R4-13	F2-08	K6-05
F2-11	D1-11	S6-05	K3-05	B4-07	A5-02	B4-04	S3-02	F5-10	A2-09	K6-02	D1-03
F6-07	D5-09	S4-11	K1-06	B2-09	A3-07	K5-04	F4-08	B3-08	D6-06	A1-05	S2-04
F3-05	D2-06	S1-03	K4-09	B5-04	A6-09	D2-09	K4-14	A6-07	F3-01	S1-02	B5-08
S1-14	A3-11	D5-01	B2-07	F6-05	K4-10	F5-12	D4-13	S3-09	K6-13	B1-01	A2-11
S2-07	A4-13	D6-07	B3-01	F1-13	K5-05	B1-08	S6-09	F2-01	A5-03	K3-03	D4-09
S4-02	A6-10	D2-03	B5-02	F3-02	K1-01	A5-08	B4-12	K6-04	S3-01	D1-10	F2-09
A3-05	B2-04	K4-13	S1-06	D5-10	F6-13	K3-14	F2-04	B1-10	D4-08	A5-05	S6-10
B5-11	S4-04	F6-12	A3-10	K1-10	D2-05	A6-11	B5-13	K1-12	S4-08	D2-07	F3-14
D5-02	K1-09	A3-06	F6-06	S4-07	B2-05	F1-01	D6-04	S5-11	K2-08	B3-04	A4-04
B3-05	S2-05	F4-13	A1-08	K5-06	D6-08	K1-11	F6-02	B5-14	D2-14	A3-03	S4-06
D6-03	K2-01	A4-11	F1-06	S5-12	B3-12	K2-05	F1-03	B6-11	D3-12	A4-09	S5-07
K6-06	F5-02	B4-06	D1-04	A2-07	S3-08	F4-10	D3-11	S2-03	K5-07	B6-09	A1-12
D1-05	K3-02	A5-04	F2-07	S6-12	B4-11	A1-13	B6-13	K2-07	S5-06	D3-07	F4-09
S5-10	A1-01	D3-10	B6-14	F4-11	K2-03	D4-03	K6-08	A2-12	F5-14	S3-05	B1-09

Unknown to three of the subjects, the first ten blocks of stimuli were presented a second time, following the first complete presentation of all 504 stimuli, to provide a check on criterion drift and reliability. Three longer rests of about five minutes were taken during the experiment, bringing the total time up to about an hour and a quarter.

The subjects were told to rate how much the speech was degraded. That is, the more degraded it appeared, the higher the number to be assigned. They were also told that some undegraded samples would be included, and that these provided an anchor point, and should be rated at zero degradation. No other constraints were imposed, except that they were asked to try to make their answers proportional, so that twice as large a number would be assigned to a system that sounded twice as degraded as another. The complete instructions that were given to the subjects to read are given in Appendix A.

The task thus has some of the properties of a magnitude-estimation task. In particular, the data should fulfil the requirements of a ratio-scale (Stevens, 1951). Unlike the usual magnitude estimation procedure, however, the scale is bounded. This seemed reasonable in scaling degradation of speech quality, unless one accepts that some vocoding system might actually be able to enhance speech quality over its undegraded form, which seems unlikely. Since the zero was set, and subjects were asked to respond proportionately, the first response they gave effectively

set their unit size. Their idiosyncratic unit size would clearly be expected to drift, over the first few dozen stimuli, while the subjects established their frame of reference. This was one reason for repeating the first 120 stimulus presentations at the end, so that the responses collected while this process was going on could be discarded.

c. Analysis of the Rating Data

We have not completed our analysis of the rating data, but the preliminary results we have obtained are encouraging. The first task in analyzing the data is to subject it to the same type of analysis as we applied to the rank-order data, to see whether we obtain comparable results. The way we have chosen to use MDPREF (the multidimensional scaling program that fits a vector-model to the data) is to pool data across subjects, and look for the effects of the different sentences in one analysis, and of different talkers in a second analysis. Pooling the ranking data across subjects was relatively simple, since each subject was constrained to use the ranks 1-14 on the 14 systems, and therefore each subject's data were directly comparable with those of the other subjects. This is not true in the rating data, where subjects were unconstrained as to the numbers they assigned, except for the anchor point for the undegraded speech. The subjects differed widely in how they chose to assign their ratings, with one subject (S1) following a logarithmic distribution (as expected if the subject followed the instructions accurately), and another (S4) following a rectangular

distribution. Histograms of subjects response distributions are shown in Figure 1. In order that the pooled ratings would not be swamped by the ratings assigned by the subject who used the largest values, each subject's raw ratings were divided by the arithmetic mean of all his 504 ratings, and the ratings of different subjects were then averaged within comparable conditions.

The resulting data were analyzed by MDPREF, and the results are plotted in Figures 2 and 3. Figure 2 shows a 2-dimensional analysis of the effects of sentence materials, and Figure 3 shows the effects of different talkers. For comparison, the results of parallel analyses of the rank-order data from the same subjects are presented in Figures 4 and 5: it can be seen that the general distribution of systems in the 2-dimensional space, and of the vectors, is highly similar for Figures 2 and 4 (effect of sentences) and in Figures 3 and 5 (effect of talkers). We have not formalized this similarity yet, but it is obvious from the figures that the agreement is good. This similarity strongly suggests that the two sets of judgments were based on a single underlying psychological structure, as we had hoped.

d. Reliability of the Data

The data collected on blocks 1-10, of twelve stimuli each can be compared with the corresponding data for the second presentation of the same stimuli, in blocks 43 through 52, for each subject. There are two aspects of this comparison we should consider: (1) Do

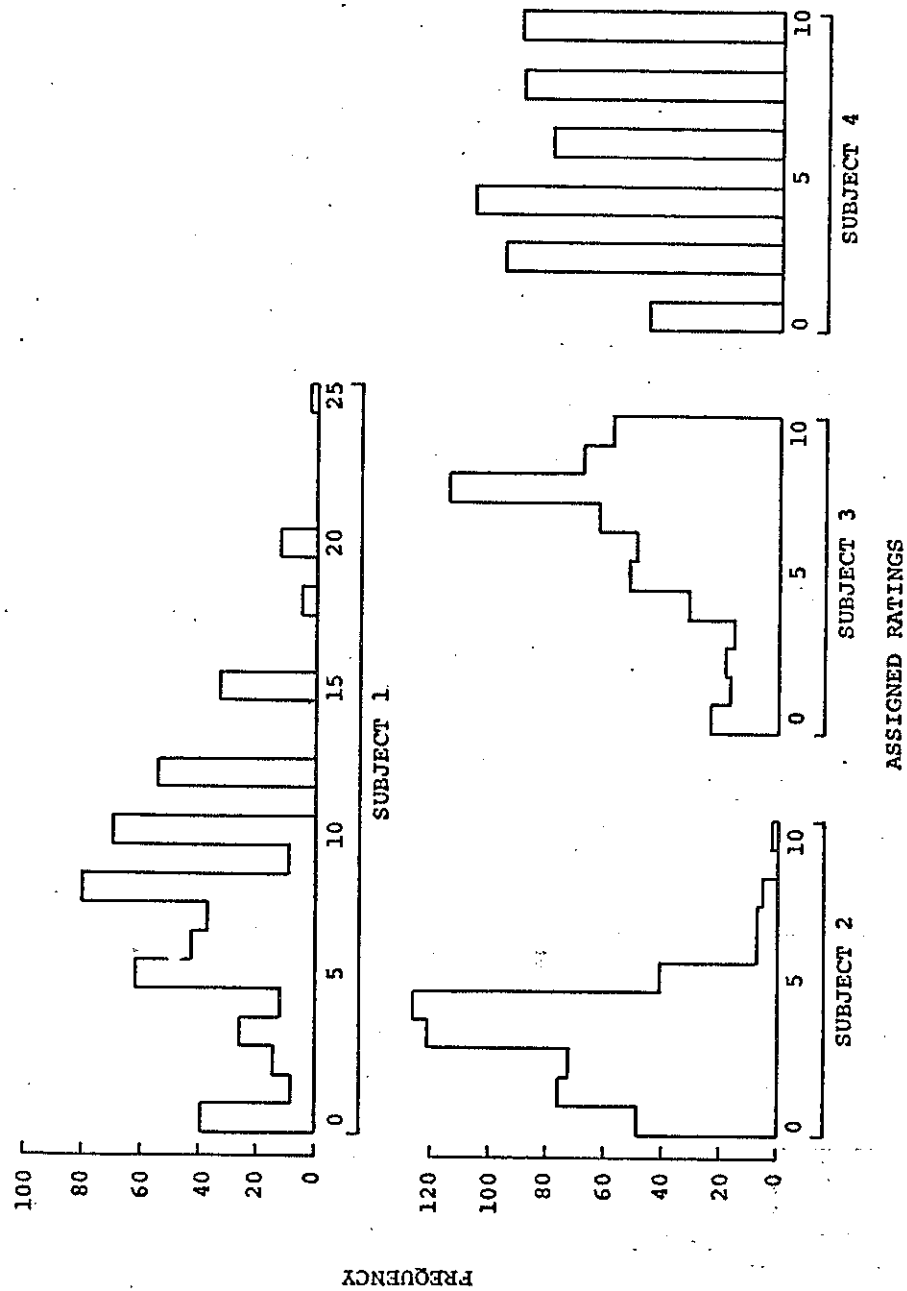


Figure 1. Distribution of ratings used by each subject over last 504 stimuli (i.e., omitting initial presentation of repeated stimuli).

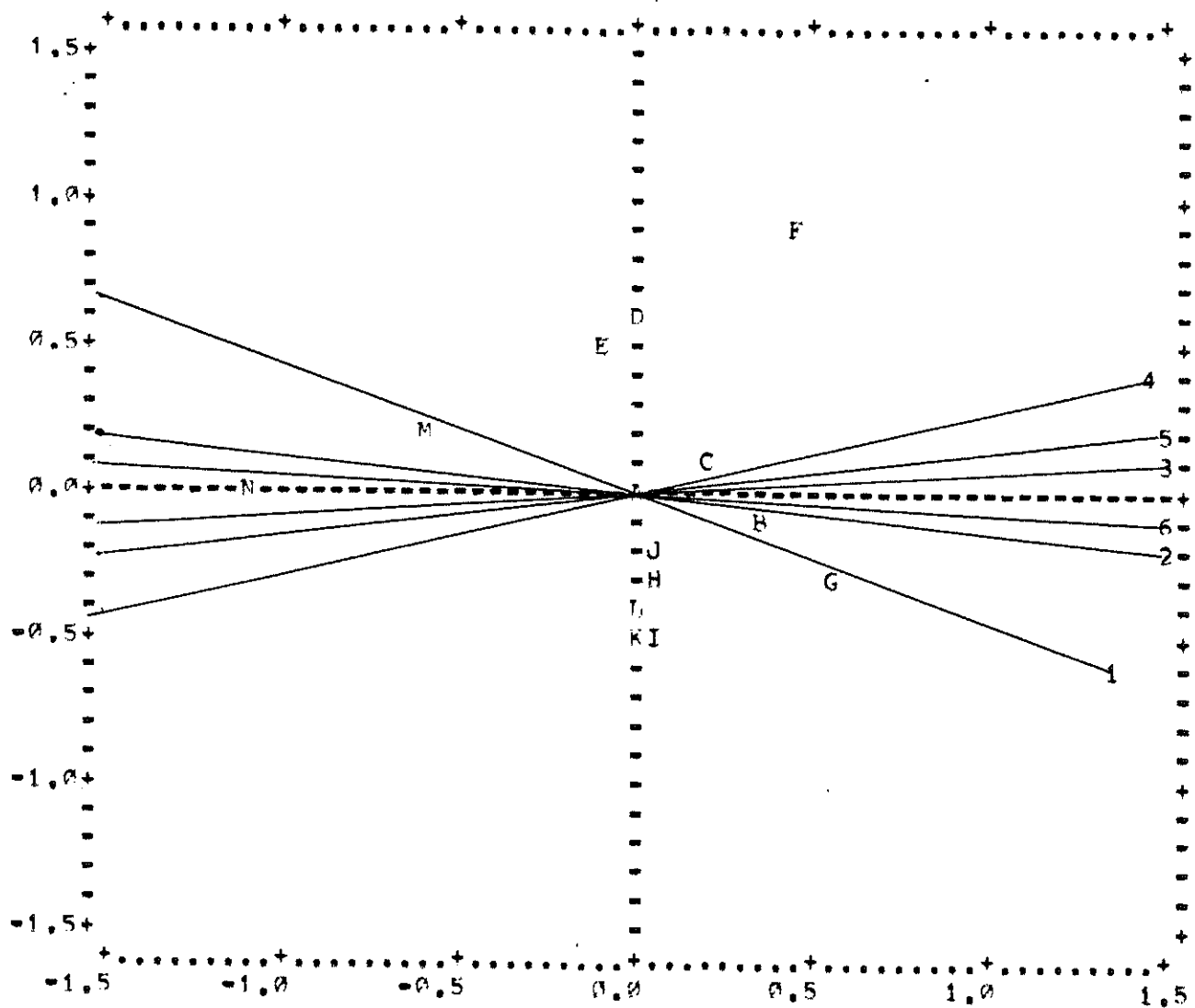


Figure 2. Two dimensional MDPREF analysis of rating data, pooled across subjects and talkers.

Systems 1-14 are represented by the letters A-N, and sentences 1-6 by the vectors labeled 1-6. Quality decreases from left to right along each of the vectors.

(System 14, letter N, corresponds to the undegraded speech.)

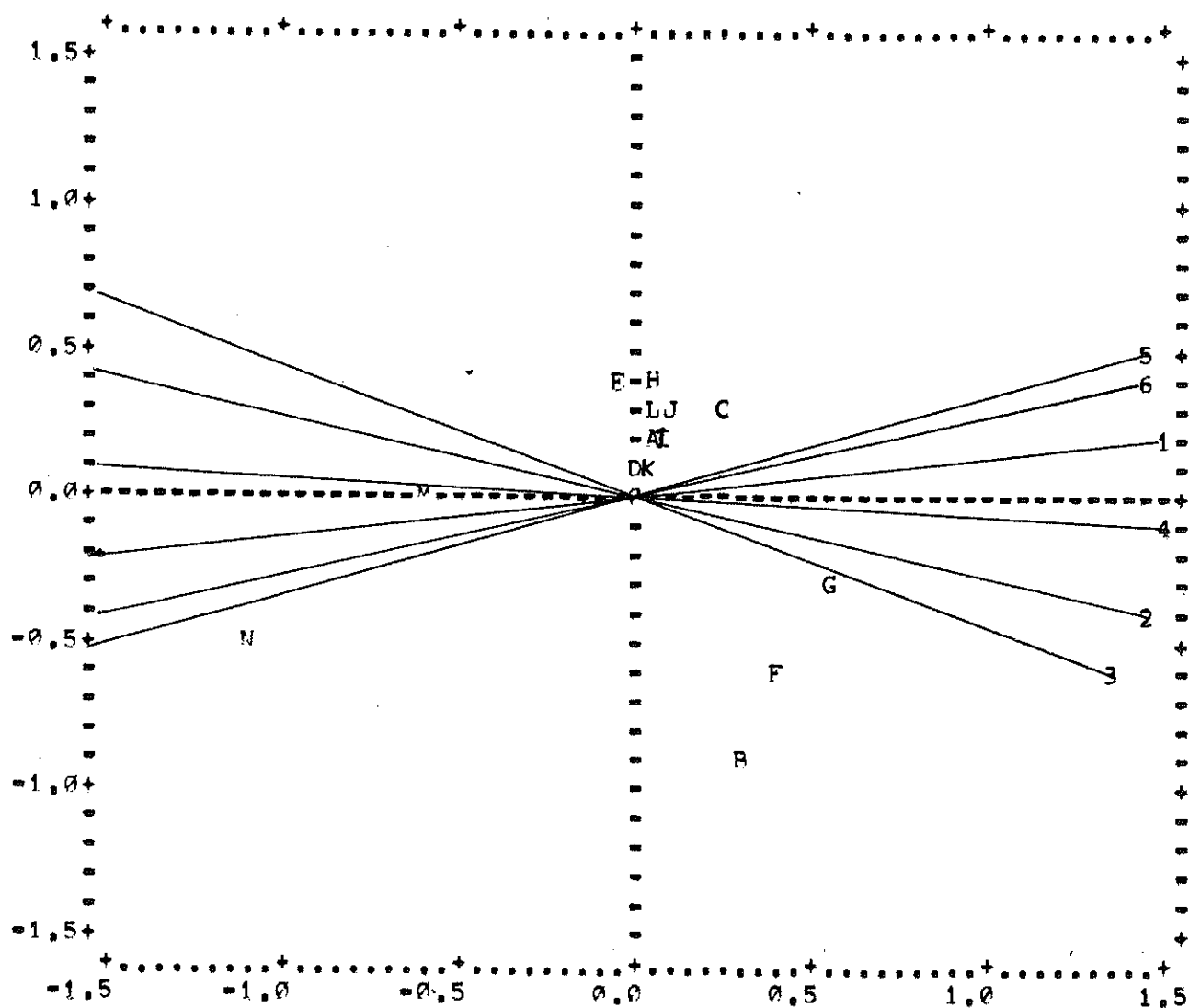


Figure 3. Two dimensional MDPREF analysis of rating data, pooled across subjects and sentences.

Systems 1-14 are represented by the letters A-N, and talkers 1-6 by the vectors labeled 1-6. Quality decreases from left to right along each of the vectors.

(System 14, letter N, corresponds to the undegraded speech.)

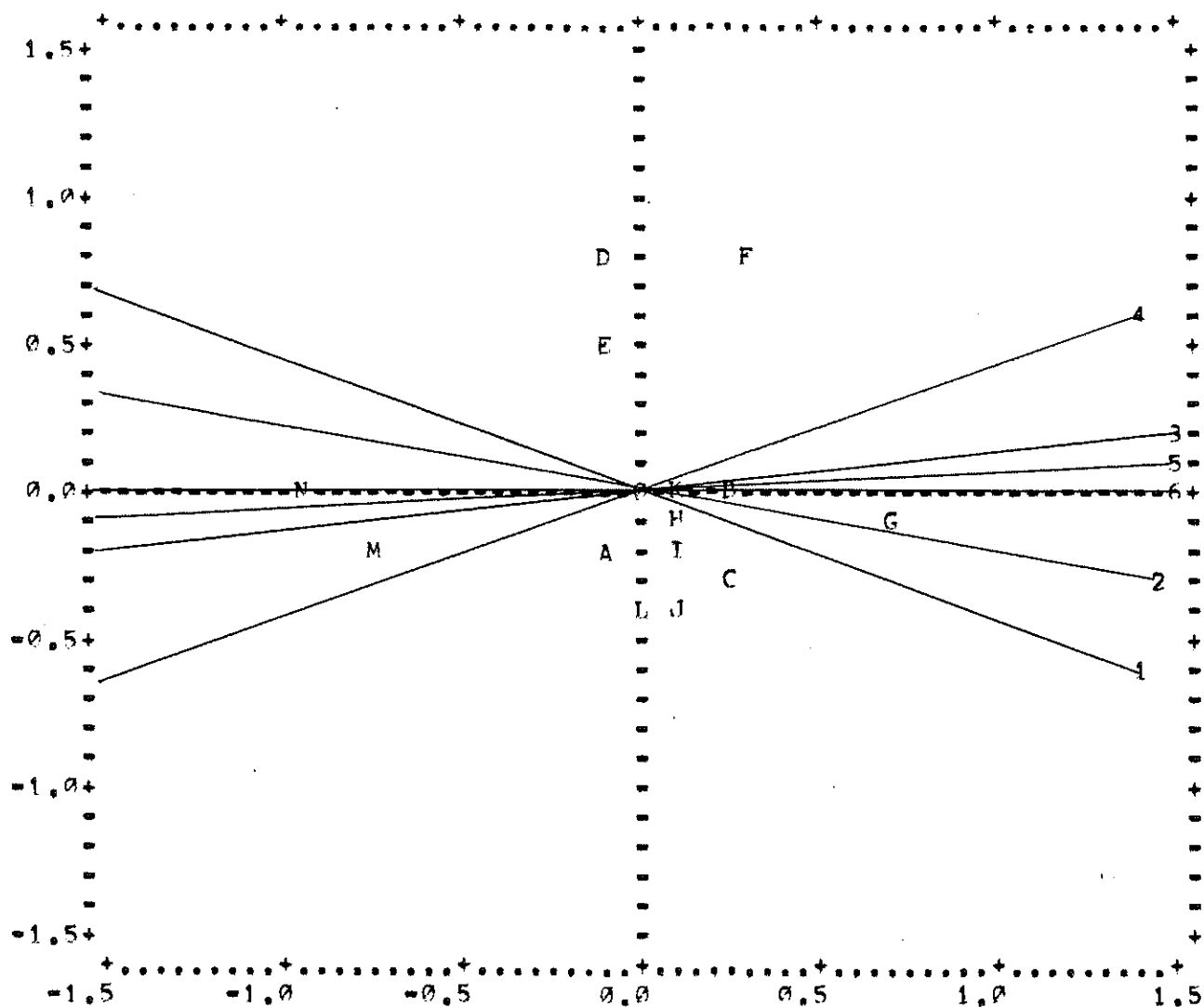


Figure 4. Two dimensional MDPREF analysis of ranking data, pooled across subjects and talkers.

Systems 1-14 are represented by the letters A-N, and sentences 1-6 by the vectors labeled 1-6. Quality decreases from left to right along each of the vectors.

(System 14, letter N, corresponds to the undegraded speech.)

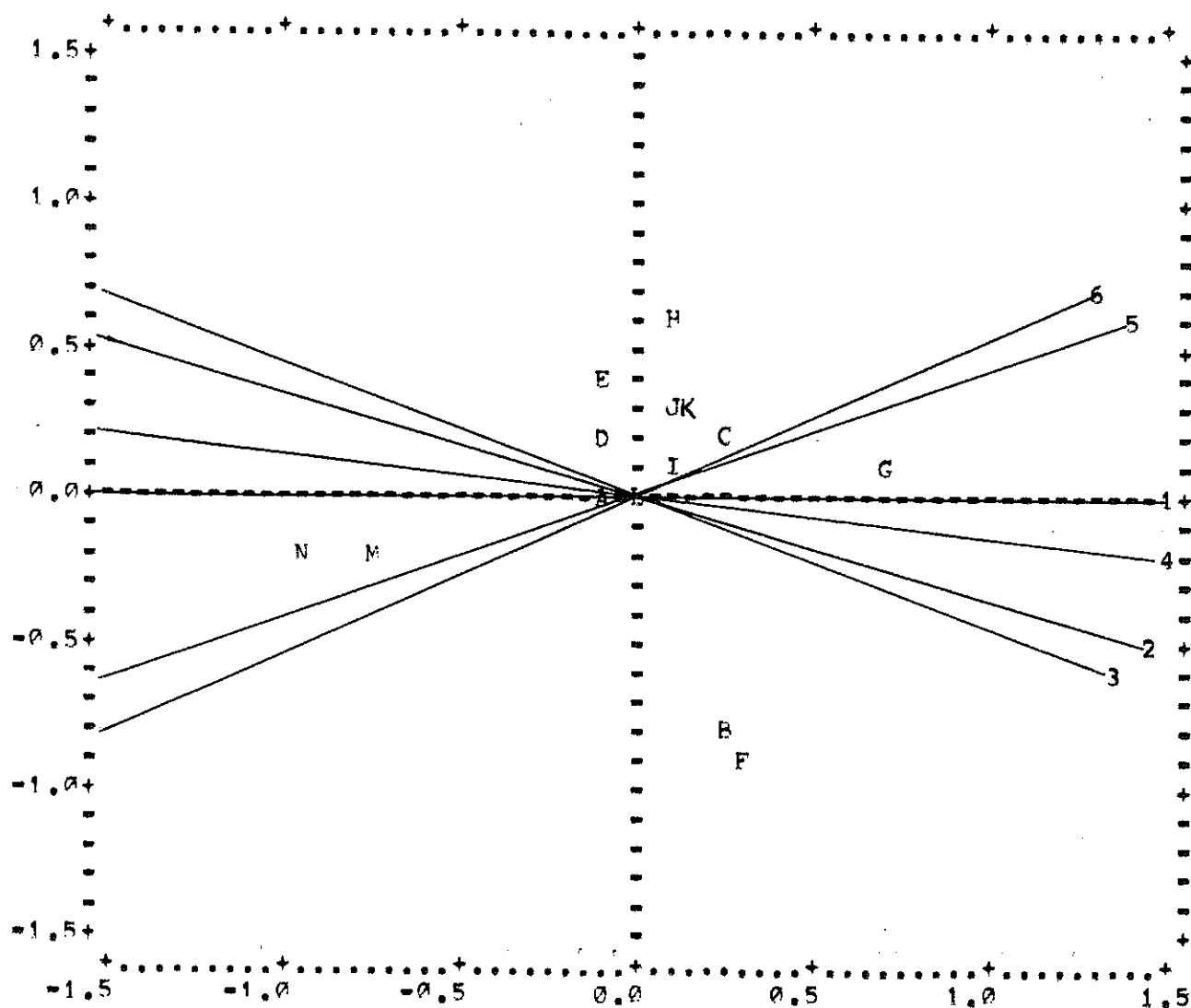


Figure 5. Two dimensional MDPREF analysis of ranking data, pooled across subjects and sentences.

Systems 1-14 are represented by the letters A-N,
and talkers 1-6 by the vectors labeled 1-6.
Quality decreases from left to right along each
of the vectors.

(System 14, letter N, corresponds to the undegraded speech.)

the data for first and second presentations correlate highly? A high correlation implies that the reliability of the data is good. (2) Does the regression line fitted to the data have a slope close to 1? A positive answer would show that the subject was using the same frame of reference for assigning his ratings at the end of the experiment as at the beginning.

The repeated-block data for each subject was split into two halves, and product moment correlations and least-square regression lines were calculated over each set of sixty pairs of observations. The correlation coefficients ranged between 0.892 and 0.975, and all are highly significant (a coefficient of 0.41 is significant at the 0.1% level of significance). Two of the subjects showed small improvements in correlation from the first to the second sets of five repeated blocks, and the other two showed no change. The correlation coefficients are plotted in Figure 6A. We can conclude that the data were highly reliable for each of the four subjects.

In Figure 6B, the regression lines are plotted for each of the four subjects, for the first and for the second sets of repeated blocks. All eight regression lines have a slope close to 1.0, and in three cases out of the four, the slope for the second set of repeated blocks (blocks 6-10) is closer to unity than the slope for the first set (blocks 1-5). This provides convincing evidence that the subjects managed to maintain the same reference system for assigning their ratings throughout the 624 item experiment (504 stimuli, with 120 repeated).

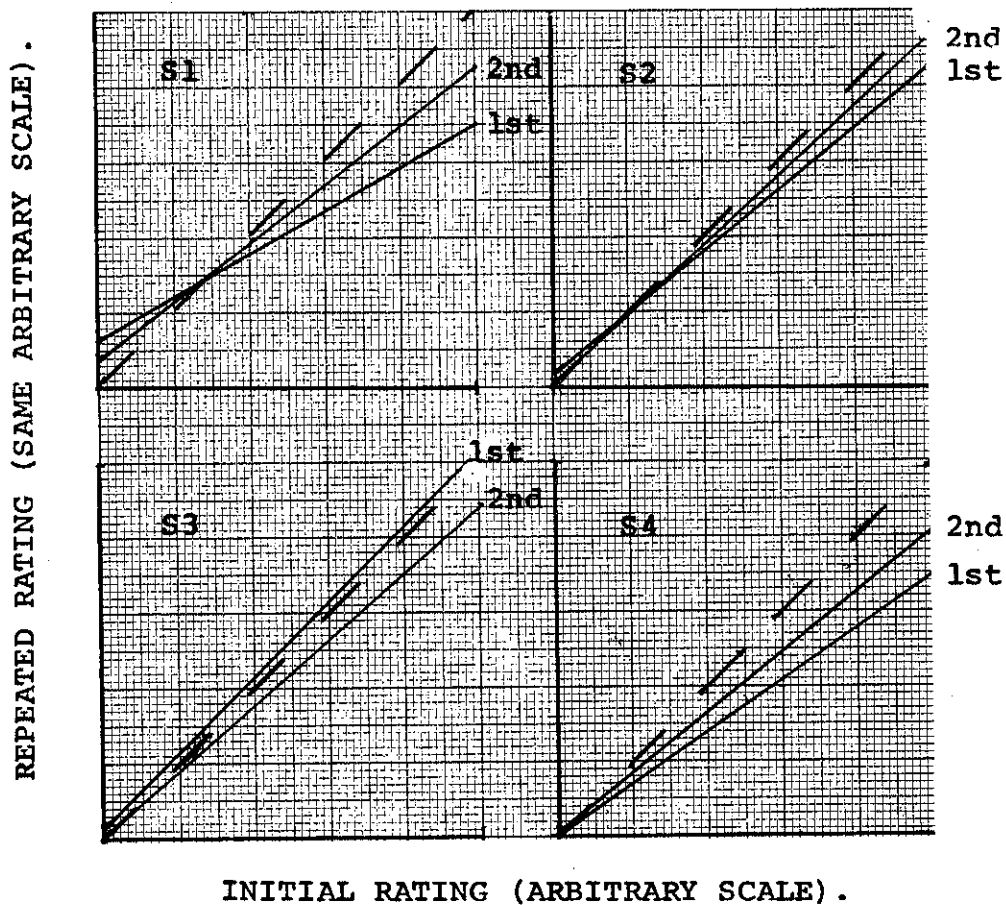
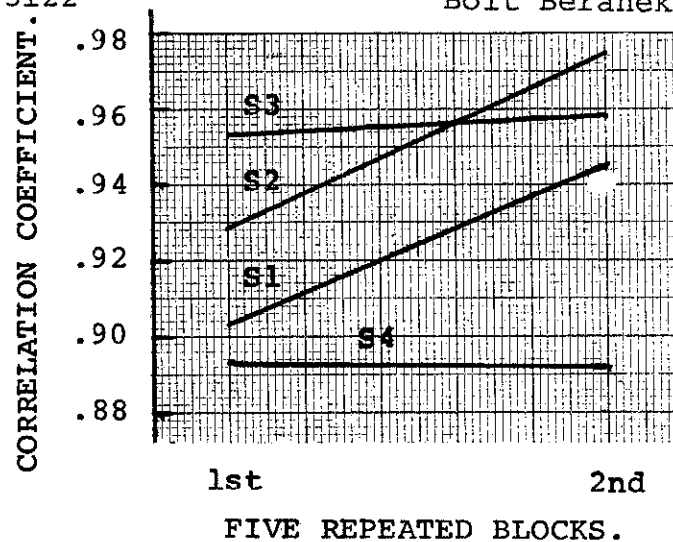


Figure 6. A: Correlation coefficients, and B: regression lines, relating repeated ratings to initial ratings for the first five (1st), and the second five (2nd) repeated blocks, for four subjects (S1-S4).

3. MULTIDIMENSIONAL SCALING PROGRAMS

We have made several minor changes and improvements to the MDS programs INDSCAL and MDPREF, supplied to us by Bell Telephone Laboratories. Some of these changes have been directed at reducing the cost of running the programs, and others improve the utility of the programs by improving the routines that produce automatic plots of the solution.

4. PHONEME SPECIFIC TESTS

We have made two complete sets of recordings for the phoneme-specific tests, as described in QPR 1, Section 7.5. The large amount of material would require a major investment of computer time to digitize and process even through a single vocoder system. Therefore, we have decided to select three or four of the 32 test lists recorded by each speaker, and use these for pilot tests with two or three systems, chosen so as to bracket the range of system qualities to which we eventually hope to apply the procedure. We have also been waiting for some improvements to be made to the real-time system we use for digitizing the speech. These improvements will be particularly useful and efficient in digitizing the phoneme-specific materials, since they include automatic editing of the waveform files. The improvements are now complete, and we expect to process the subset of materials in the next few weeks.

5. PAPERS

We have begun preparing a rather comprehensive review of subjective techniques that have been used to evaluate the quality of speech transmission systems. The intent is to produce a paper that will critically review the various approaches that have been taken to the problem of speech quality assessment. We are describing the procedures in some detail, and attempting to point out their advantages and limitations.

Among the methods that will be covered are: the isopreference method first used by Munson and Karlin (1962), the relative preference method introduced by Hecker and Williams (1965), the absolute preference or rating scale method (Pachl, Urbanek & Rothauser, 1971), the category judgment method (Grether & Stroh, 1973), the rank-ordering method (Nakatani, 1972), the scaling of individual speech qualities (Coolidge & Reir, 1959; Stevens, Nickerson, Rollins, & Boothroyd, 1974), and multidimensional scaling techniques (Carroll, 1972). Each of these techniques has been used by several investigators and the report will cover, insofar as possible, the work of all of them.

The paper also addresses the problem of controlling for talker, sentence material, and listener effects in the results of evaluation studies. Also, the possibility of utilization of objective measures of quality is considered, and recommended procedures for carrying out effective quality evaluations are described. These recommended procedures are based, both on the review of previous work and on the

results obtained to date in the current project. This paper will be finished during the next quarter and will be submitted as part of the final report.

Another paper describing some of our work to date will be presented at the upcoming meeting of the Acoustical Society of America. The paper, entitled Some Effects of Speech Materials on Vocoder Quality Evaluations, will be presented by A.W.F. Huggins.

6. FUTURE PLANS

- a. Further analysis of the ranking data and the rating data. This will include analysis with INDSCAL, and correlations between the solutions for the ranking and rating data. In addition, we will test the extent to which the ranking data are included in the rating data: the ratings assigned to the 14 systems will be rank-ordered for each of the 36 speaker-sentence combinations, and these derived rank orders will be compared with the rank orders obtained empirically in the ranking task. The most important aim of this further analysis is to establish, if possible, what are the perceptual dimensions along which the speech quality of the various systems vary.
- b. If the aim specified in (a) is achieved, we will attempt to scale the systems on these dimensions directly. This represents an extension of the work

described in Section 3 of QPR 2.

- c. Perform the phoneme specific pilot-tests, as outlined in Section 4 of the present report.
- d. Make additional recordings under conditions appropriate for real-life use of the ARPA vocoder, process the materials through the systems developed by the ARPA contractors, and apply our testing procedures to these processed recordings.

REFERENCES

- Carroll, J. D., "Individual Differences and Multidimensional Scaling", in R. N. Shepard, A. K. Romney, & S. Nerlove (Eds.), Multidimensional Scaling: Theory and Applications in the Behavioral Sciences. Vol. 1 Theory. New York: Seminar Press, pp. 105-155, 1972.
- Coolidge, O. H., & Reir, G. C., "An Appraisal of Received Telephone Speech Volume", Bell System Technical Journal, Vol. 38, p. 877, 1959.
- Grether, C. B. & Stroh, R. W., "Subjective Evaluation of Differential Pulse-Code Modulation Using the Speech 'Goodness' Rating Scale", IEEE Transactions on Audio & Electroacoustics, Vol. AU-21, pp. 179-184, 1973.
- Hecker, M. H. & Williams, C. E., "On the Interrelation Among Speech Quality, Intelligibility, and Speaker Identifiability", Fifth Congress International d'Acoustique, 1965.
- Huggins, A. W. F., "The Perception of Timing in Natural Speech: I. Compensation Within the Syllable", Language and Speech, Vol. 11, pp. 1-11, 1968.
- Munson, W. A. & Karlin, J. E., "Isopreference Method for Evaluating Speech-Transmission Circuits", Journal of the Acoustical Society of America Vol. 34, pp. 762-774, 1962.
- Nakatani, L. H., "Rank Ordering Procedure for Auditory Stimuli", J. Acoust. Soc. Amer., Vol. 51, pp. 1370-1372, 1972.
- Pachl, W. P., Urbanek, G. E., & Rothauser, E. H., "Preference Evaluation of a Large Set of Vcoded Speech Signals", IEEE Transactions on Audio and Electroacoustics, Vol. AU-19, pp. 216-224, 1971.
- Stevens, K. N., Nickerson, R. S., Rollins, A., & Boothroyd, A., "Use of a Visual Display of Nasalization to Facilitate Training of Velar Control for Deaf Speakers", BBN Report No. 2899, September, 1974.
- Stevens, S. S., "Mathematics, Measurement, and Psychophysics", in S. S. Stevens, (Ed.) Handbook of Experimental Psychology. New York, Wiley, pp. 1-49, 1951.

Appendix A

INSTRUCTIONS FOR QUALITY DEGRADATION RATINGS

We have developed a Voice-Coding system for improving the efficiency of transmitting speech, for example over telephone lines. The price paid for improving the efficiency is that the Vocoder degrades the quality of the speech. The purpose of this experiment is to find out how much degradation is introduced by each of several different Vocoders.

We have recorded six speakers, each reading the same six sentences. The recordings have then been processed through several Vocoders. You will hear a total of 624 recorded sentences, and your task is to assign a number to each sentence you hear, corresponding to the amount you think the speech has been degraded. Included in the sequence, along with the processed versions, are some of the original recordings which have not been degraded at all. Assign the number 0 to the undegraded sentences. The more degraded a sentence is, the higher the number you should assign to it. Try to assign numbers proportionately - that is, if one sentence is twice as degraded as another, assign a number twice as big. Don't worry about making mistakes during the first dozen or two, but try to be consistent after that.

The sentences are recorded in blocks of 12. Write your responses against the corresponding numbers, and use a new page for each block of twelve sentences. Don't change earlier judgments, and

don't look back at earlier pages. If you need to rest, say so in
the gap between two blocks of twelve, and the experimenter will stop
the tape.